3D Tracking of Deformable Objects with Applications to Coding and Recognition

Juan Ruiz-Alzola, Carlos Alberola-López*, J.R.Casar-Corredera**,Gonzalo de Miguel-Vela**
EUIT Telecommunication-ULPGC, 35017 Las Palmas, Spain
*ETSI Telecommunication-UVA, 47011 Valladolid, Spain
**ETSI Telecommunication-UPM, 28040 Madrid, Spain
Tel : +34-28-452862. Fax : +34-28-451243
e-mail : jruiz@cibeles.teleco.ulpgc.es

ABSTRACT

In this contribution we address the problem of motion and structure estimation of objects that fit a deformation model. Our purpose is to provide a suitable input to a recognition system directed at detecting particular shapes and deformation patterns (gestures) of the object. This is accomplished by means of a stereoscopic vision system which first reconstructs 3D tokens -points- from the images; then the tokens are tracked independently in order to obtain an improved estimation of their positions and to keep a correspondence among them in consecutive instants of time. Finally the tokens are matched to an allowed state -shape- of a Finite State Machine which depicts the deformation of the body. Rigid motion is considered to relate the actual tokens positions with the estimated shape. This approach provides with a convenient way to deal with incomplete collections of measurements due to occlusions.

1.- INTRODUCTION

Motion and structure estimation of 3D objects is a key issue in computer vision and related fields. Although rigid objects have received a great deal of attention for many researchers [1] leading to many algorithms and results, this has not been the case with deformable (non-rigid) objects. Nevertheless, most objects in our world possess some kind of non-rigidity and modern computer vision systems must take this circumstance into account. In this contribution we address the problem of motion and structure estimation of non-rigid objects when some kind of *a priori* information about the objects is known. Our purpose is to use the estimated information of structure and motion in coding both the deformation and global motion of the body under study. This coding should be easy to use as the input to a recognition system intended to command several tasks. One promising approach to this problem is to establish a parametric model for the different shapes of the 3D object[2] : for each different shape there is a one to one mapping to a minimal set of parameters satisfying that smooth variations of the shape should lead to smooth variations of the parameters; we will call *configuration* to a particular set of parameters. A sensing device obtains a collection of measurements -geometric primitives- related in a functional way to the parametric model ; then the model is fitted to these measurements by means of global optimization techniques. A general model for the measurement equation could be :

$$\mathbf{z}[n] = \mathbf{H}[n]\left(\mathbf{R}[n]\mathbf{f}\left(\boldsymbol{\alpha}[n]\right) + \mathbf{t}[n]\right) + \mathbf{w}[n] \qquad \textbf{(1)}$$

where n is the current discrete-time, $\boldsymbol{\alpha}$ is the parameters vector, **f** is a vector valued function of the parameter set which translates the configuration into the measured coordinates (usually cartesian) referred to a reference system intrinsic to the body, **R** is a rotation matrix and **t** is a translation vector which represent the motion of the shape from a standard position to the actual position in the space, **w** is a noise vector due to the measurement process and **H** is the measurement matrix which selects the tokens that are actually measured (for example, if the tokens were points some of them are occluded by the body and they cannot be measured). Therefore, our concern is to obtain an estimation of vector $\boldsymbol{\alpha}$ (structure estimation) and matrix **R** and vector **t** (motion estimation) from the measurements **z** in each instant of time n.

There are some drawbacks in this approach : First the optimization loop does not guarantee not to get trapped in false minima, second the convergence can be very slow making difficult a real-time implementation, third occlusions lead to incomplete descriptions of the shape and it is not easy to control which possible solution the algorithm will converge to. These drawbacks can be somehow alleviated considering the smoothness condition stated previously and performing a tracking of the parameters through time; then, at each instant of time, the optimization process is fed by the predicted parameters. In this paper we try to provide an

alternative framework to this problem with the aim of obtaining a real-time recognition system suitable for many kinds of objects.

2.- A MODEL-BASED STEREO FRAMEWORK TO ESTIMATE STRUCTURE AND MOTION

Two cameras operating in a stereoscopic manner are used as sensors with the aim of simplifying the solution of eq. 1 using 3D data instead of 2D data from monoscopic images. We call *tokens* to the geometric primitives of the object which are measured. The selected collection of tokens is useful to depict the configuration of the body only if we can get the parameters vector α from them. For example, articulated bodies can be considered formed by links and joints; a configuration is specified by angles between links and the tokens can be the joints and the free extremes of the limbs. It is clear that α can be obtained from the cartesian coordinates of the tokens; nevertheless, due to occlusions on the images some tokens are lost and the rest of them can satisfy many configurations of the body. Note that if the system is monoscopic the measured coordinates are related to the 3D tokens by non-linear perspective projections which besides loose the depth information.

We propose to model the coherence of the deformation by a Finite State Machine (FSM), mapping each structure or configuration of the body onto a state of the FSM. For each state there are only some allowed transitions to other configurations. The FSM represents a quantization of all possible trajectories in the configurations (parameters) space. It is possible to obtain a coarse approximation -useful for some applications- of the deformation with just a few states, being necessary many states for an accurate one. There are several reasons to use a FSM: first, given a set of measured tokens we can obtain only valid configurations of the model (those corresponding to allowed transitions from the current state) ; second, it is feasible to incorporate physical constraints to the model just pruning some transitions; third, coding the deformation by a chain of states is a natural framework to many recognition tasks; fourth, the ambiguity on the current configuration due to occlusions is reduced to a few allowed transitions of the FSM being possible to develop feasible ad-hoc approaches to the problem; fifth, the paradigm can be extended to more general automata (stochastic, fuzzy) which can improve the matching of collections of tokens to states and the later recognition process.

Herein we present an approach to estimate the current state of the non-rigid body based on tracking some outstanding tokens on the body surface :

2.1.- The Proposed Algorithm

The algorithm can be split into the following steps :

1) Select some outstanding patches of the body, extract them from each image and compute their centroids; these are our measured tokens and they must be labeled. As it will be seen this process is guided by temporal tracking. Note that tokens extraction is a noisy process.

2) Reconstruct the 3D cartesian coordinates of the tokens. A main problem in stereo-vision is that of the correspondence among tokens in both images; in our case this problem is overcome by the labeling of the measured tokens. To label the 2D tokens we consider the temporal coherence of their motion and other properties such as color, texture, etc.

3) A tracking filter can provide improved estimations of the 3D tokens positions. There has been an intensive use of Kalman Filters in Vision to solve this question [4]. The optimality of these filters rely on some assumptions which are not always easy to meet. As a feasible alternative in many circumstances we propose to use simpler constant-coefficients trackers such as the well-known $\alpha\beta$ filters [5]. These filters operate uncoupledly on each 3D token cartesian coordinate and independently on different tokens and assume as kinematical model a deterministic constant velocity motion. The filter equations (for each coordinate) are :

$$r_{is}[n] = r_{ip}[n] + \alpha\left[z_i[n] - r_{ip}[n]\right] \qquad \textbf{(2.a)}$$

$$v_{is}[n] = v_{ip}[n] + \frac{\beta}{T}\left[z_i[n] - r_{ip}[n]\right] \qquad \textbf{(2.b)}$$

where n is the current discrete time, i stands for each token number (label), s for smoothed (filtered), p for predicted for current time in the previous one, r is the cartesian coordinate (x, y or z), v is the cartesian velocity coordinate (v_x, v_y or v_z) and z is the measured cartesian 3D position coordinate. T is the sampling period and α, β are the filter coefficients. These coefficients can be chosen by means of an array of criteria being very application dependent; for example [6] trades-off noise smoothing and transient capability. It must be noticed when designing these filters that noise statistics depend on token positions in 3D space and are different for each coordinate.

4) The estimated positions of the tokens must be matched to a state of the FSM. This problem can be

complicated due to noise and occlusions and it will be consider later.

5) We consider a cartesian reference system intrinsic to the object. Each configuration of the body has a unique representation of its tokens in this reference system; in other words, each still configuration of the body can be regarded as a rigid body and the mentioned reference system is rigidly attached to it. Global motion is estimated by fitting this "rigid body" to the set of 3D measured tokens [3,7].

6) Predict next discrete time position and velocities of the tokens. We use a constant velocity motion model for each point :

$$r_{ip}[n+1] = r_{is}[n] + v_{is}[n]T \qquad \textbf{(3.a)}$$

$$v_{ip}[n+1] = v_{is}[n] \qquad \textbf{(3.b)}$$

8) Project the predicted 3D positions onto both cameras. The next time we will look for tokens inside correlation gates around the predicted 2D positions. Time correspondence among measurements can be readily accomplished by a nearest neighbor algorithm operating on the tokens inside the correlation gates and considering the similarity of other properties (as color) too. Other more elaborated algorithms can be also employed [5]. Now we go to step one again.

2.2.- Matching Measurements To The States.

Suppose that we have got a body described by N_T tokens and by a FSM with N_S states. Due to occlusions we can obtain only N_M measurements ($N_M<N_T$) in the current instant of time; moreover, the body was in state S_0 in the previous instant of time and according to the FSM it is evolving to any of the states S_0 S_1 ... S_p. As it has been previously mentioned, each state of the FSM is defined by a parameter vector α or, alternatively, by the positions of the whole set of tokens in an intrinsic reference system. In this section the collection of the current measurements is referred to this reference system.

An observation vector is formed stacking the current measurements. In the absence of occlusions the vector is $(\mathbf{z}_1,...\mathbf{z}_{NT})^t$. The effect of occlusions can be modeled by multiplicating this vector by matrix **H** which extracts the actually measured tokens. The squared distance between two complete set of observations can be defined as :

$$d^2(\mathbf{z},\mathbf{z}') = (\mathbf{z}-\mathbf{z}')^t \mathbf{D}(\mathbf{z}-\mathbf{z}') \qquad \textbf{(3)}$$

where **D** is a symmetric, positive-definite weighting matrix. If we had two incomplete set of observations **Hz** and **H'z'** the distance evaluation is constrained to the common measurements.

To decide to which state the current observation vector (possibly incomplete) belongs we compute the distance between the ideal positions of the complete set of tokens of each possible configuration and the current observation vector. If any of those distances drop below a determined threshold the corresponding state is matched. Note that due to occlusions several states can be satisfied. These states could be incompatible among them (in the sense that each different selection would lead to a different deformation of the part of the body which is seen) and then the decision should be deferred until new observations throw more light on the subject. A different situation arises when the ambiguities lead to different possible shapes in the occluded side of the body; in this case it is better not to choose a determined state but to consider that any configuration is possible where we are not seeing what is going on.

3.- EXPERIMENTS

Now we present an application of the previous ideas to the problem of estimating the global motion and sequence of states of a computer simulated hand.

The hand is formed for the plane of the palm and four common fingers and a thumb that stem from it. Each of the four fingers are characterized by the tip, two interfalangeal joints with one degree of freedom (extension/flexion) and a joint connecting the palm and the finger with two degrees of freedom (ext/flex and abduction). The flexion/ extension motion is performed in planes orthogonal to the palm and by self-inspection it can be seen that the angles for both interfalangeal joints use to be similar. The abduction separate the fingers and defines the flexion plane. When totally extended these fingers are in the plane of the palm. Notice that each finger can be considered as a planar articulated stick moving in its flexion plane and that this plane is defined to be orthogonal to the palm and containing the tip and the palm joint of the finger. The thumb has one interfalangeal joint (ext/flex), one finger-palm joint (ext/flex) and a joint near the wrist and not visible with two degrees of freedom (ext/flex and abduction). For each abduction there is a flexion plane for the thumb which is not orthogonal to the palm; moreover, when the thumb is extended it is not contained in the plane of the palm (note that the thumb place against the other fingers). It is possible to define an intrinsic reference system for the hand with axes **X** in the plane of the palm, axes **Y**

orthogonal and going towards the inside of the palm and **Z** = **X**x**Y**. We consider that the hand only has global or deformation motion at a time.

In each instant of time we have measurements of the positions of the tips of the five fingers and of the joint linking the palm to the thumb. We assume that we also have measurements of some fixed points on the palm which determine its plane and the intrinsic reference system. We consider that the joints linking the four common fingers to the palm remain fixed in the palm and therefore we know their position in each instant of time.

The flexion of each finger is modeled by a FSM whose states are defined by the quantized angles at the three joints (0, π/6, π/3, π/2 for each joint). Transitions are allowed only for adjacent states. The four common fingers only need two parameters to represent their state because two joint angles use to be similar. The abduction is discretized to two possible values (0, π/4) corresponding to two flexion planes.

Several continuous deformation patterns have been simulated corresponding to some gestures of the human hand. The previously mentioned measurements have been corrupted with gaussian white noise. These measurements are reconstructed, filtered and referred to the intrinsic reference system. Then the plane of the palm is fitted to the measurements and the abductions angles are obtained projecting the tips of the fingers on this plane; for the thumb this is achieved from the measured joint. Finally, a flexion state is matched for each finger.

As an example, figure 1 shows the possible positions of the tips of the first finger (index) in its flexion plane for each of the sixteen states. The horizontal axes is contained in the palm plane and the palm-finger joint is in (0,0). We consider the hand making a fist in a time second and a sampling frequency of 25 frames/sec. The solid line links the states corresponding to this deformation and the dotted lines represent possible transitions from these states. Figure 2 shows the real temporal sequence of states (solid line) and the detected sequence for a typical case with a noise with standard deviation of 10 pixels (dotted line).

4.-CONCLUSION

In this paper we have presented a general method to estimate the structure and motion of non-rigid objects which are able to fit a deformation model. This method is intended to facilitate the subsequent task of shape and/or deformation pattern recognition in real-time. Current work is oriented towards the extension of the FSM to more involved paradigms which are able to incorporate physical constraints and preferences on the deformation patterns in an easier fashion. Major interest is placed in the development of articulated models to code hand gestures and gait.

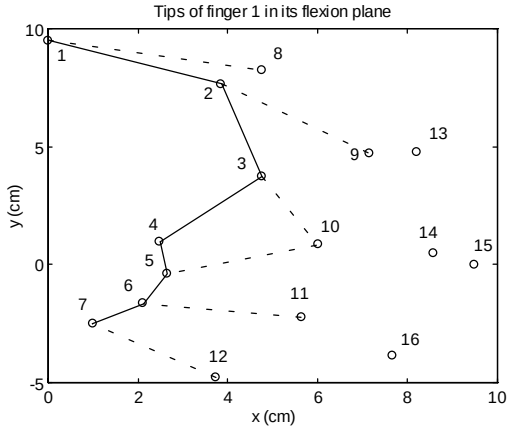

Figure 1

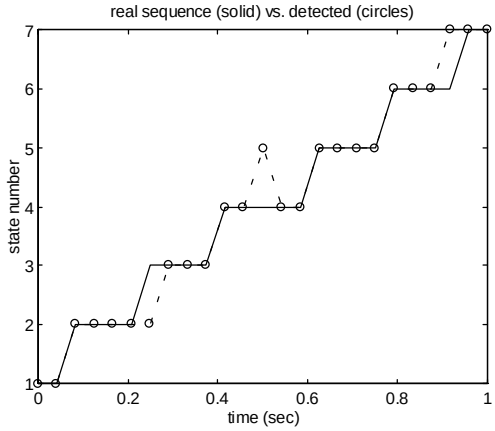

Figure 2

5.-REFERENCES

[1] T.S.Huang,A.N.Netravali, "Motion and Structure from Feature Correspondences: A Review", Proc.of the IEEE, vol.82,No.2,February 1994,pp.251-268.

[2] D.G.Lowe, "Fitting parameterized 3D models to images", IEEE Trans.PAMI,vol.13,No.5,1991,pp.441-450.

[3] K.S.Arun,T.S.Huang,S.D.Blostein, "Least-Squares Fitting of Two 3-D Point Sets" ,IEEE Trans.PAMI, vol.9, No.5,September 1987,pp.698-700.

[4] O.Faugeras, "Three-Dimensional Computer Vision", MIT Press, Cambridge, MA 1993.

[5] S.S.Blackman, "Multiple Target Tracking with Radar Applications", Artech-House, MA 1986.

[6] T.R.Benedict,G.W.Bordner, "Synthesis of an Optimal Set of Radar Track-While-Scan Smoothing Equations", IRE Trans.on Automatic Control, vol.AC-7, No.4, July 1962, pp.27-32.

[7] J.Ruiz-Alzola, C.Alberola, J.R.Casar, "Rigid Object Tracking from Long Sequences of Stereoscopic Images: A Recursive Constant Coefficients Filter Approach", SIP,IASTED, Las Vegas, NE, nov.1995.